DOCUMENT RESUME

ED 469 163 TM 034 453

AUTHOR	Sykes, Robert C.; Hou, Liling; Hanson, Brad; Wang, Zhen
TITLE	Multidimensionality and the Equating of a Mixed-Format Math Examination.
PUB DATE	2002-04-00
NOTE	23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 2-4, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Elementary School Students; *Equated Scores; Error of Measurement; Intermediate Grades; Item Response Theory; *Mathematics Tests; Test Construction; *Test Format
IDENTIFIERS	Anchor Effects; *Multidimensionality (Tests)

ABSTRACT

This study investigated the effect on student scores of using anchor sets that differed in dimensionality in item response theory (IRT) scaled tests. Real data from a mathematics achievement test that had been documented to have dimensions aligned with item format were used. Item responses were available from a representative sample of approximately 2,600 fifth graders taking a mathematics test from a large state assessment. The use of anchor set items that differed in the degree their items loaded on each of the two salient factors resulted in varying amounts of equating effort. The equatings using anchors that contained items loading more heavily on the first or the second dimension had standard errors (square root of total error) of 6.46 and 3.44, representing 13% and 7% respectively of the approximate 50 scale score standard deviation. The standard error of equating anchors with items that were balanced in terms of their loadings on the two dimensions was 2.72, 5%, of a standard deviation. Bias constituted a substantial portion of the total error of the equatings using the "unbalanced" F1 and F2 anchor sets. (Contains 7 tables, 6 figures, and 17 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 469 163

TM

Multidimensionality and the Equating of a Mixed-format Math Examination

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
R. Sykes
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

Robert C. Sykes

Liling Hou

Brad Hanson

Zhen Wang

CTB/McGraw-Hill

This paper was presented at the Annual Meeting of the National Council on Measurement in Education in New Orleans, April 2002.

TM034453

BEST COPY AVAILABLE

INTRODUCTION

The use of both c.r. and m.c. item formats in forms equated through common item equating requires determining their presence in anchor item sets. Accepted practice calls for the set of common items to be proportionally representative of the total test forms in content and statistical characteristics (Kolen & Brennan, 1995). The belief that the anchor item set should constitute a miniature version of the test is substantiated by research that has documented that inadequate common item content representation can impact test scores when examinee groups taking alternate forms differ considerably in achievement level (Klein and Jarjoura, 1985). One possible explanation of how inadequate anchor set content can impact equating is the presence of other dimensions underlying performance or the insufficient control of these dimensions. Beguin, Hanson, and Glas (2000) in a simulation study noted a large effect of multidimensionality on IRT equating for nonequivalent groups.

Utilizing anchor item sets consisting exclusively of m.c. items is a frequent practice, however. This is done presumably only when there are not any significant content categories that are represented only by c.r. items or when there is evidence that a single dimension adequately explains the item responses. The use of exclusively m.c. anchor item sets offer the substantial advantage of allowing equating to take place in the frequently narrow time frames demanded by the rapid turn-around of scores as it doesn't require conducting a rater drift study to monitor and

correct, if necessary, for systematic changes in rater judgments on the c.r. items over time.

Some authors have contended that the use of exclusively m.c. anchor sets can serve to correct for these differences between different set of readers of c.r. responses across test administrations (Ercikan et al, 1998). However, Tate (2000) has noted through simulations the potential for serious bias in the estimation of IRT linking constants with the use of m.c. only anchor item sets when unidimensionality is violated and when multidimensionality is induced by systematic changes in rater judgments over time.

The purpose of this research was to investigate the effect on student scores of using anchor sets that differed in dimensionality in IRT-scaled tests. Real data from a Math achievement test that had been documented to have dimensions aligned with item format was utilized.

METHOD

Source Data

Item responses were available from a representative sample of approximately 2,600 fifth grade students taking a Math field test of a large state assessment. The content of the items of the field test form - 73 m.c. and 16 short answer (two-point) c.r. items for a total of 105 points - was representative of the state curriculum standards. Only students who responded to at least 2/3's of the items were used. Omits were treated as not

correct.

A set of 45 items (10 c.r.) selected on the basis of the test blueprint and thus representative of the state curriculum standards had been selected from this field test form to constitute the first operational form. Item responses were available for large samples of fifth graders randomly selected from the operational testing population.

Analyses

Verification of Dimensionality

A previous study of item responses from the operational exam using Poly-Dimtest (Li & Stout, 1995) had demonstrated the presence of at least two dimensions. A different trait was measured by the exclusively c.r. items constituting the AT1 subtest than that measured by the remaining m.c. items. Another sample of students (S1) was drawn from the operational administration to replicate the results from the previous study.

The presence of a similar multidimensionality associated with item type in the field test form (pool) was assessed through Poly-Dimtest analyses. Eigenvalues from factor analyses of the item responses were examined to determine whether a similar number of dimensions significantly explained performance on the field and operational tests.

Construction of Anchor Item Sets

Four modifications or variants of the operational form, each containing the same items but with different sets of those items designated as anchors, were constructed. The four anchor sets

were constructed to be content representative and to approximate the difficulty of the operational test.

Two of the anchor sets (B1 and B2) were baseline anchor sets in that the loadings of the items on the significant dimensions were balanced, having similar average loadings. Two other anchor sets (F1 and F2) contained items that loaded more heavily on one of each of the significant dimensions.

Evaluation of Forms

The origin of the items of the operational form in the field test pool permitted equating the form with each of the four anchor sets to the field test scale using the Stocking-Lord procedure (1983). The equating was used to place the item parameter estimates for each form/anchor set onto the field test scale.

Number-correct scale score estimates (described below) were then obtained for a second large sample of students taking the operational form using the four sets of item parameter estimates. Arbitrarily assigning the equating of the first (B1) of the two balanced anchor item sets as the criterion, the discrepancy measure of Petersen, Cook, & Stocking (1983) was used to evaluate the three other equatings.

The discrepancy measure:

$$\frac{\sum_j f_j d_j^2}{n} = \frac{\sum_j f_j (d_j - \bar{d})^2}{n} + \bar{d}^2, \qquad (1)$$

is a weighted mean square difference (WMSD) that is the sum of the variance of the difference and the squared bias. In the

first two terms of the equation, $d_j = (t'_j - t_j)$ where t'_j is the estimated scale score from the compared equated form (B2, F1, and F2) for raw score x_j and t_j is the criterion (B1) scale score for x'_j. f_j is the frequency of x_j and $n = \sum_j f_j$ is the sample size. The third term is $\bar{d} = \frac{\sum_j f_j d_j}{n}$.

Rating Process

Readers were trained to implement scoring rubrics; anchor papers, check sets, and read behinds were employed to verify and maintain scoring accuracy. Inter-rater reliability studies that incorporated second reads for a large sample of students taking each test indicated that the percentage of exact agreement on the c.r. items in the field test ranged between 93.01% and 100.00%. Exact agreement rates for the two-point c.r. items of the operational test ranged between 89.90% and 97.56%.

Scaling Process

Multiple-choice and open-ended items were scaled together using the generalized IRT model. With the generalized model a three-parameter logistic model (Lord, 1980) was used for the multiple-choice items:

$$P_i = P(X_i = 1|\theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7 A_i(\theta - B_i)]}, \qquad (2)$$

where A_i is the discrimination, B_i is the difficulty, and c_i is

the lower asymptote or guessing parameter for item i.

A generalization of Master's (1982) Partial Credit model was used for the c.r. items. This two-parameter partial credit (2PPC) model is the same as Muraki's (1992) "generalized partial credit model." For a c.r. item with m_i score levels assigned integer scores that ranged from 0 to m_i - 1:

$$P_{ik}(\theta) = P(X_i = k-1|\theta) = \frac{\exp(y_{ik})}{\sum_{j=1}^{m_i}\exp(y_{ij})}, \qquad k = 1,\ldots,m_i \qquad (3)$$

where

$$y_{ik} = \alpha_i(k-1)\theta - \sum_{j=0}^{k-1}\gamma_{ij},$$

and $\gamma_{i0} = 0$. α_i is the item discrimination. γ_{ij} is related to the difficulty of the item levels: the trace lines for adjacent score levels intersect at γ_{ij}/α_i.

Parameter Estimation

Item parameter estimation was conducted using the program PARDUX (Burket, 1991; 1995). Item parameters were estimated for the field test form and each of the four operational forms/anchor sets using marginal maximum likelihood procedures implemented with an EM algorithm. Evaluations of the accuracy of the program with simulated data (Fitzpatrick, 1990) have found it to be at least as accurate as MULTILOG (Thissen, 1986).

The ability scale was defined by specifying a prior true θ distribution to have a mean of 0.0 and standard deviation of 1.0 for the field test sample. Field test item parameter estimates

were linearly transformed to a scale score metric having a mean of approximately 500 and standard deviation of approximately 50. The LOSS and HOSS (lowest and highest obtainable scale scores) were set for the field test form at 250 and 850. For the four equated form/anchor sets, a LOSS of 275 and HOSS of 750 were applied.

Student Scores

The relationship between the (predicted) raw score and estimated scale score (tcc) was obtained using the final item parameter estimates:

$$E(X_a \mid \hat{\theta}_a) = \{\sum_{i=1}^{mc} P_i(\hat{\theta}_a) + \sum_{j=1}^{cr} \sum_{k=1}^{m_j} (k-1) P_{jk}(\hat{\theta}_a)\}, \qquad (4)$$

where the predicted raw score has been partitioned into components for the *mc* multiple choice items and the *cr* constructed response items.

RESULTS

Raw score descriptive statistics for the field test and operational form are presented in Table 1. The field test was more difficult than the subset of items chosen as the operational form. The mean raw score of 38.71 was less than 40% of the total number of points (102) compared to the greater than 50% average performance reflected by the average score of 30.93 out of the total 55 points for the operational form.

The difference in performance on the field test relative to the operational form reflects both the presence of some more difficult items in the field test, especially the presence of some more difficult c.r. items, as well as differences in student

performance. The average of the p-values (average score divided by total points) of the 16 c.r. items in the field test was .14. The average p-value for those 10 c.r. items selected for the operational test was .19 in the field test administration compared to .38 in the operational administration. The mean p-value for the operational m.c. items also increased, from .52 in the field test versus .66 in the operational administration(not shown). The overall .15 average p-value increase for the operational items likely arises from differences in student motivation and perhaps from possible differences in the ability of the student populations tested on the two occasions.

Dimensionality Assessments

Table 2 contains the Poly-Dimtest (Li & Stout, 1995) significance tests of the hypothesis of unidimensionality. Both the field test and operational forms were multidimensional (p-values less than .00) using exclusively c.r. items in the AT1 subtest.

The eigenvalues from a principal factor analysis of the reduced product moment correlation matrix obtained from the SAS FACTOR procedure (SAS, 1988), using squared multiple correlations as prior communality estimates, are presented in the Scree plots for the field test in Figure 1 and the operational test in Figure 2. The factor analysis of product moment correlations may result in the presence of spurious difficulty factors. Unfortunately the test length and sample size precluded the employment of a more appropriate exploratory item factor analysis employing

tetrachoric correlations for the m.c. items and polychoric correlations for the c.r. items. The first factor for both tests is very large relative to all others. Only the first two eigenvalues for each test are greater than 1.0.

Differences between the eigenvalues, presented in Table 3 for the six largest eigenvalues, are less than .2 starting with the third and fourth eigenvalues for each test (.131 and .175 for the field test and operational form, respectively). The third and higher factors explain less than 6% of the common factor variance if the effect of negative eigenvalues for the later rejected factors is accounted for.

Form/Anchor Set Comparisons

Anchor Item Sets

Field test descriptive statistics for the items selected for the operational form and the four anchor sets are presented in Table 4. Each of the anchor sets was constructed to be content representative in that the number of items in each of the significant content categories was within 10% of that called for by the state curriculum standards (i.e. test blueprint). Each anchor set consisted of 12 items and between 13 and 15 points (one to three c.r. items in the F2 and B1 anchors, respectively). There were between zero (F1 and F2) and six items (B2 and F2) in common among the six possible pairs of the four anchor sets.

Average anchor set item difficulty as represented in the field test statistics was very similar to that for the set of operational items, ranging between .42 (F1) and .50 (F2) versus

.45 for the operational form.

The average loading of the items of each anchor set on the first and second factor of the operational administration are provided in Table 5. These factors were the first two principle factors obliquely rotated. After the Promax rotation the two factors correlated .58 with each *uniquely* accounting for 2.779 and 2.308 of the 9.082 common factor variance (sum of the first two eigenvalues of Table 2) after eliminating other factors. The average F1 and F2 loading for the B1 and B2 anchor sets (.24 versus .24 and .23 versus .24, respectively) are more similar than the average F1 and F2 loading for the operational test (.25 versus .21 in Table 1).

Loadings of the items on both of the factors are presented in the Appendix, first for the c.r. items then the m.c. items. It is worth noting that while almost all the c.r. items load more heavily on the first factor (and the sixth c.r. item is only a marginal exception) there are a number of m.c. items that load more heavily on the first factor. Thus while the second factor may be characterized as a "m.c. factor" because only m.c. items load more appreciably on it, the first factor cannot be similarly characterized as a c.r. factor.

Equatings

Test characteristic curves (tccs) for the four Stocking-Lord (1983) equatings are presented in Figures 3 - 6. The alignment of the tccs for the equated anchor sets with their input values (field test) was very good for the two baseline anchor sets (B1

and B2) and for F2. The equating using the F1 anchor set resulted in a greater deviation between the anchor tccs.

The total error or WMSD and the component sources of error are given in Table 6. Relative to the comparison of the two baseline equatings, the F1 anchor equating has more than five times total error (41.7540 versus 7.4162), with more than 95% of the error attributed to bias (40.0179). The F2 equating also produced substantially greater total error (11.8544) than the equating of the baseline anchors with 37% of the error in the form of (squared) bias (4.3551). The comparison of the baseline B2 equating with the criterion B1 equating indicated a very small amount of squared bias (.5798), amounting to 8% of the relatively small amount of total error (7.4162).

The first four moments of the scale score distributions obtained from a second large, representative sample obtained after the Stocking-Lord transformation constants were used to place the four sets of operational item parameter estimates onto the field test scale are provided in Table 7. The two baseline anchor sets produce mean scale scores that differ the least: .77 (544.40 - 543.63). The mean scale score for the distribution produced from the F2 equating was more than two points greater than those produced with the baseline anchor sets. The F1 mean demonstrates that the large bias present with the equating of the F1 anchor set results in the underestimation of scores. The mean for the distribution produced from F1 is less than the means for the two baseline anchors by more than five points.

DISCUSSION AND CONCLUSIONS

The employment of anchor item sets that differed in the degree their items loaded on each of the two salient factors resulted in varying amounts of equating error. The equatings employing anchors that contained items loading more heavily on the first or the second dimension had standard errors (square root of total error) of 6.46 and 3.44, representing 13% and 7% (respectively) of the approximate 50 scale score standard deviation. The standard error of equating anchors with items that were balanced in terms of their loadings on the two dimensions was 2.72 or 5% of a standard deviation.

Bias constituted a substantial portion of the total error of the equatings using the "unbalanced" F1 and F2 anchor sets, between 13% and 4% of a standard deviation. Tate (2000) also found a bias in the estimation of IRT linking coefficients using an extension of the Stocking-Lord procedure for the graded response model. In a simulation of a bidimensional test where one dimension measured a c.r. ability and the other a m.c. ability that correlated .6, Tate noted that anchor sets that were unbalanced with respect to item type (i.e. exclusively m.c. items) underestimated the simulated increase in abilities relative to anchor sets that were balanced across item type. The exclusively m.c. anchors failed to capture the large change in the mean of the c.r. ability.

The increase in scores on both the m.c. and c.r. items was large between the field and operational tests of this study. The magnitude and breadth of the increase in the scores on the c.r. items, as well as the use of validity check sets and other rater reliability procedures that produced high rater agreement rates, suggests increased student performance as opposed to an increase in the leniency of the population of raters. Given the absence of evidence of a cohort effect and the diminished pattern of item omissions between field and operational testing, the increases are more likely substantially attributed to increased motivation on the operational test.

On the IRT trait constituting a composite of two, item- type related dimensions correlating .58, mean performance increased by approximately 40 scale score points (80% of a standard deviation). The direction of the biases resulting from the use of the F1 and F2 anchors appear to reflect differential performance of the groups taking the pilot and operational forms on the two dimensions. A multiple-group confirmatory item factor analysis that tests whether the factor patterns are the same across administrations and estimates latent means has been planned. Constraints in the form of the size of the test will be dealt with by estimating the appropriate correlation matrix and inputting it to a structural equation program such as EQS (Bentler & Wu,1995) that can handle larger tests/samples. Results can then inform a simulation that models the particular dimensional structure of the Math test and allows comparisons of

results against true multidimensional abilities.

The presence of equating bias should serve as a cautionary note about constructing anchor sets that are not miniature tests. Ensuring that an anchor set is a miniature test and thus representative of the significant facets of performance requires knowledge of the dimensional structure of the exam.

The identification of a significant facet of performance associated with c.r. items also points to the potential for trends in rater judgment to confound population changes in ability. The implementation of rater trend studies that assess change over time by scoring previously scored papers that are seeded in the operational scoring stream would appear to be a necessary step for testing programs that utilize c.r. items to make significant decisions about the examinees. Test users should be cognizant of the need for these studies and allow scoring windows of sufficient length to accommodate them.

REFERENCES

Beguin, A.A., Hanson, B.A., & Glas, C.A. (2000). *Effect of Multidimensionality on separate and concurrent estimation in IRT equating.* Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

Burket, G.R. (1991; 1995). *PARDUX.* Monterey, CA: CTB/McGraw-Hill.

Bentler, P.M., & Wu, E.J.C. (1998). *EQS for Windows (Version 5.7)*[Computer software]. Encino, CA: Multivariate Software, Inc.

Ercikan, K, Schwarz, R., Julian, M.W., Burket, G.R. Weber, M.W., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35,* 137-154.

Fitzpatrick, A.R. (1990). Status report on the results of preliminary analyses of dichotomous and multi-level items using the PARMATE (PARDUX) program. Unpublished manuscript.

Joreskog, K. & Sorbom, D. (1993). *LISREL 8: Structural modeling with the SIMPLIS command language.* Chicago, Il: Scientific Software International.

Klein, L.W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197-206.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices.* New York: Springer Verlag.

Li, H., & Stout, W. (1995). A version of Dimtest to assess latent trait unidimensionality for mixed polytomous and dichotomous item response data. Paper presented at the annual conference of the National Council of Measurement in Education, San Francisco.

Lord, F.L. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum associates.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Petersen, N.S., Cook, L.L., & Stocking, M.L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8,* 137-156.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

SAS (1988). *SAS/STAT User's Guide*. Cary, NC; SAS Institute, Inc.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37,* 329-346.

Thissen, D. (1986). *MULTILOG: Multiple categorical item analysis and test scoring, Version 5.* Mooresville, IN: Scientific Software.

Table 1
Raw Score Descriptive Statistics for the Field Test and Operational Form

Statistic	Field Test (Total Points = 102)	Operational Test (Field Test Stats) (Total Points = 45)	Operational Test (S1) (Total Points = 55)
Mean	38.71	21.89	30.93
St. Dev.	14.52	8.84	10.25
Mean p-value[1]	0.41	0.45	0.60
St. Dev. p-value	0.22	0.22	0.20
Mean C.R. p-value[1]	0.14	0.19	0.38
St. Dev. C.R. p-value	0.14	0.14	0.19
Mean 1st Factor Load	0.21	0.26	0.25
St. Dev. 1st Factor Load	0.19	0.16	0.18
Mean 2nd Factor Load	0.15	0.15	0.21
St. Dev. 2nd Factor Load	0.16	0.16	0.18
Feldt-Raju reliability	0.92	0.87	0.90
N	2379	2379	2762

[1] The average score divided by the total number of points

Table 2
Poly-Dimtest Significance Tests for the Hypothesis of Unidimensionality

Test	Number of Items	T	p-value
Field Test	89	3.11	.00
Operational	45	3.44	.00

Table 3
First Six Eigenvalues from the Principal Factor Analyses

	Field Test			Operational Test		
Number	Eigenvalue	Difference	Proportion[1]	Eigenvalue	Difference	Proportion[1]
1	10.191	8.234	0.61	7.969	6.856	0.86
2	1.957	0.965	0.12	1.113	0.607	0.12
3	0.992	0.131	0.06	0.506	0.175	0.05
4	0.861	0.105	0.05	0.331	0.064	0.04
5	0.756	0.040	0.05	0.267	0.031	0.03
6	0.716	-	0.04	0.236	-	0.03

[1] Sum of the proportion of variance explained exceeds 1.00 across all factors because of the presence of later negative eigenvalues

Table 4
Descriptive Statistics for the Selected Operational Items and Four Anchor Item Sets
(Field Test Responses)

Statistic	Operational Test	B1	B2	F1	F2
Mean p-value[1]	0.45	0.46	0.46	0.42	0.50
St Dev p-value	0.22	0.17	0.14	0.17	0.19
Mean item-test cor.	0.37	0.41	0.37	0.37	0.40
St Dev item-test cor.	0.08	0.09	0.08	0.09	0.06

[1] The average score divided by the total number of points

Table 5
Means and Standard Deviation of Anchor Item
Loadings on the Two Factors
(Operational Responses)

Statistic	B1	B2	F1	F2
Mean F1 Loading	0.24	0.23	0.32	0.17
St Dev F1	0.10	0.10	0.09	0.14
Mean F2 Loading	0.24	0.24	0.16	0.31
St Dev F2	0.10	0.12	0.14	0.14

Table 6
Equating Discrepancy Measure

Anchor Sets Compared	WMSD	Var. of Difference	Squared Bias
B2 vs B1	7.4162	6.8364	0.5798
F1 vs B1	41.7540	1.7361	40.0179
F2 vs B1	11.8544[1]	7.4994	4.3551

[1] WMSD does not equal the sum of the Variance of Difference and Squared Bias due to rounding

Table 7
Scale Score Descriptive Statistics for the Field Test and Equated Operational Form

	Field Test	Operational Test Anchor Item Set (n=5525)			
Statistic	(n=2379)	B1	B2	F1	F2
Mean	498.85	544.40	543.63	538.07	546.48
Standard Deviation	53.49	51.40	49.09	52.20	49.08
Skewness	-1.31	-1.05	-1.18	-0.95	-1.22
Kurtosis	3.99	5.07	5.96	4.42	6.18

Figure 1
Scree Plot of the Eigenvalues for the Field Test

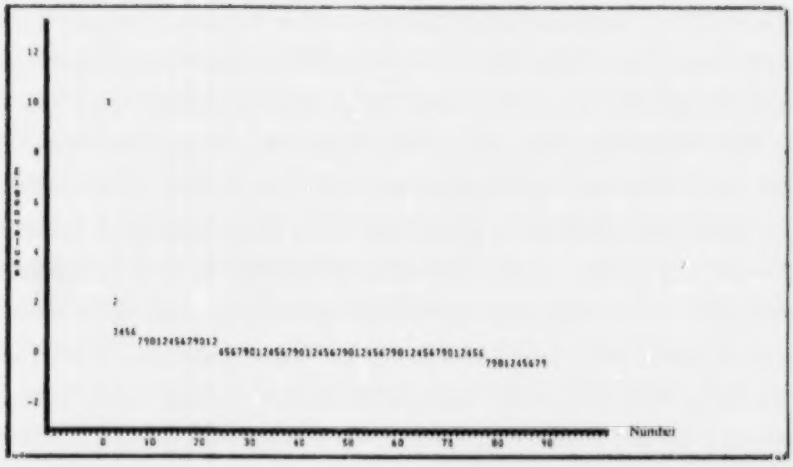


Figure 2
Scree Plot of Eigenvalues for the S2 Sample of the Operational Test

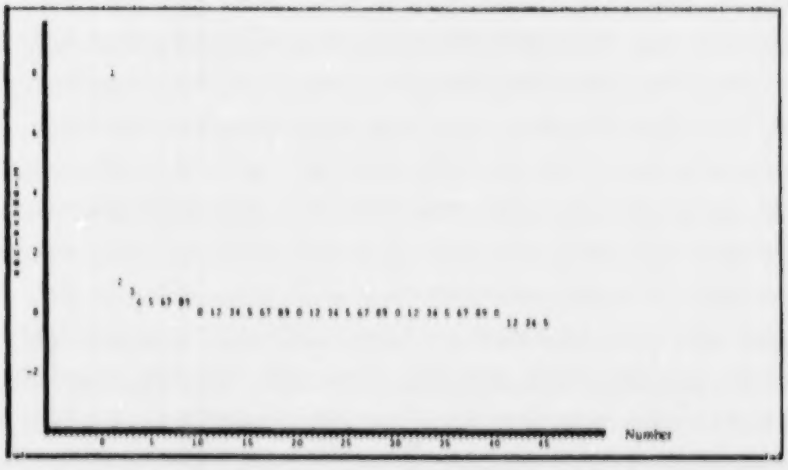


Figure 3
TCCs for the Equating using Anchor Set B1

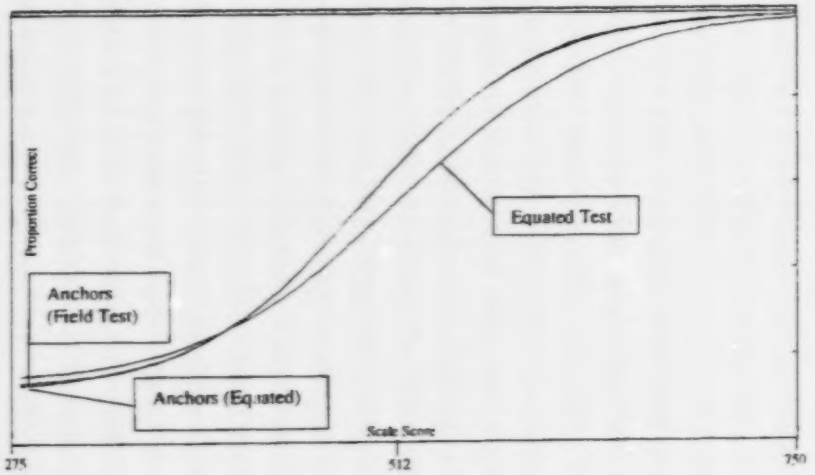


Figure 4
TCCs for the Equating using Anchor Set B2

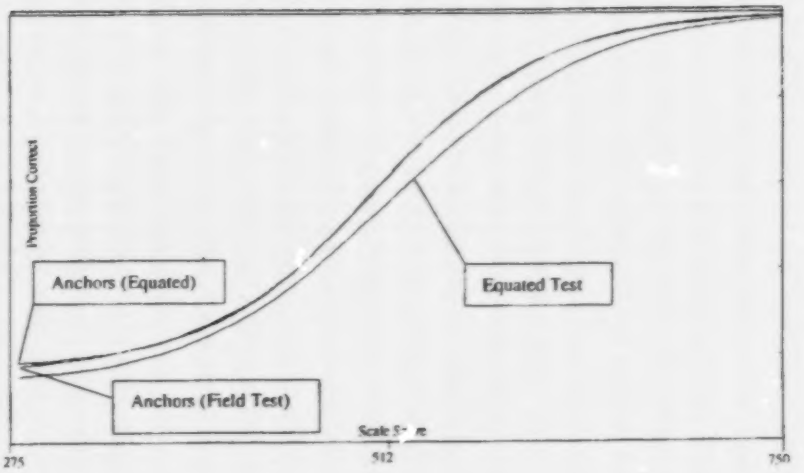


Figure 5
TCCs for the Equating using Anchor Set F1

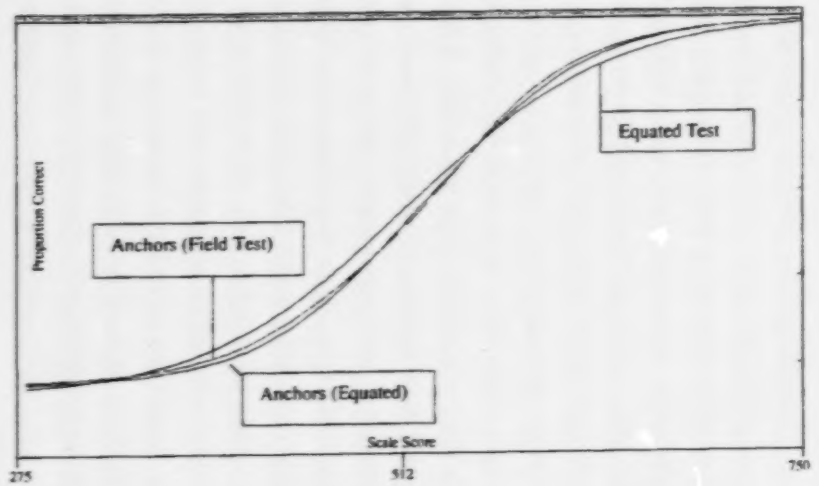
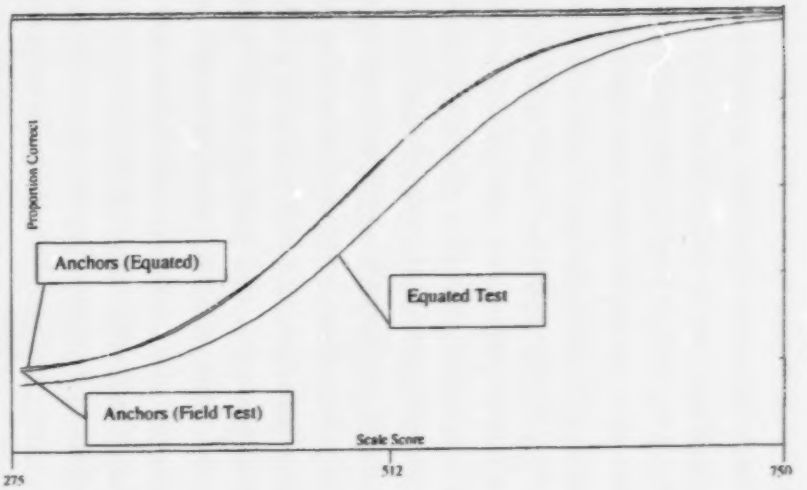


Figure 6
TCCs for the Equating using Anchor Set F2



Appendix

Rotated Factor Pattern for the Operational Administration

Item	Factor 1	Factor 2
CR.1	0.39416	0.13835
CR.2	0.57685	-0.07875
CR.3	0.64238	-0.15722
CR.4	0.33031	0.27033
CR.5	0.58793	-0.23835
CR.6	0.29857	0.29980
CR.7	0.43546	0.06186
CR.8	0.34051	0.25115
CR.9	0.49139	0.01402
CR.10	0.53603	-0.06321
MC.11	-0.13939	0.25867
MC.12	0.30377	0.11365
MC.13	0.32734	0.08266
MC.14	0.19465	0.26937
MC.15	0.26812	0.24789
MC.16	0.04234	0.42393
MC.17	0.05993	0.40734
MC.18	0.20997	0.26404
MC.19	0.16845	0.30323
MC.20	0.37501	0.10376
MC.21	0.30069	0.15853
MC.22	0.20536	0.46525
MC.23	-0.06603	0.47707
MC.24	0.26573	0.11950
MC.25	0.20874	0.16103
MC.26	-0.11934	0.54291
MC.27	0.30585	0.09054
MC.28	0.27177	0.27421
MC.29	0.24500	0.11833
MC.30	0.36558	0.01643
MC.31	0.10640	0.24018
MC.32	0.17552	0.16469
MC.33	-0.11406	0.53410
MC.34	0.18628	0.25451
MC.35	0.13807	0.32851
MC.36	0.19649	0.34130
MC.37	0.21306	0.37749
MC.38	0.39813	0.02648
MC.39	0.25597	0.33166
MC.40	0.12694	0.45505
MC.41	0.23525	0.21324
MC.42	0.40406	0.05253
MC.43	0.18976	0.36381
MC.44	0.28707	0.08841
MC.45	-0.02328	0.46667

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

TM034453

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Multidimensionality and the Equating of a Mixed-format Math Examination	
Author(s): Robert C. Sykes, Liling Hou, Brad Hanson, Zhen Wang	
Corporate Source: CTB/McGraw-Hill	Publication Date: April 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Robert C. Sykes	Printed Name/Position/Title: Research Scientist III	
Organization/Address: CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA. 93940	Telephone: (831) 393-7774	FAX:
	E-Mail Address: rsykes@ctb.com	Date: 9/6/02

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: info@ericfac.piccard.csc.com
WWW: http://ericfacility.org

EFF-088 (Rev. 2/2000)